G08RBF – NAG Fortran Library Routine Document

Note. Before using this routine, please read the Users' Note for your implementation to check the interpretation of bold italicised terms and other implementation-dependent details.

1 Purpose

G08RBF calculates the parameter estimates, score statistics and their variance-covariance matrices for the linear model using a likelihood based on the ranks of the observations when some of the observations may be right-censored.

2 Specification

```
 SUBROUTINE G08RBF(NS, NV, NSUM, Y, IP, X, NX, ICEN, GAMMA, NMAX,
1                  TOL, PARVAR, NPVAR, IRANK, ZIN, ETA, VAPVEC,
2                  PAREST, WORK, LWORK, IWA, IFAIL)
 INTEGER           NS, NV(NS), NSUM, IP, NX, ICEN(NSUM), NMAX,
1                  NPVAR, IRANK(NMAX), LWORK, IWA(4*NMAX), IFAIL
 real              Y(NSUM), X(NX,IP), GAMMA, TOL, PARVAR(NPVAR,IP),
1                  ZIN(NMAX), ETA(NMAX), VAPVEC(NMAX*(NMAX+1)/2),
2                  PAREST(4*IP+1), WORK(LWORK)
```

3 Description

Analysis of data can be made by replacing observations by their ranks. The analysis produces inference for the regression model where the location parameters of the observations, θ_i, $i = 1, 2, \ldots, n$, are related by $\theta = X\beta$. Here X is an n by p matrix of explanatory variables and β is a vector of p unknown regression parameters. The observations are replaced by their ranks and an approximation, based on Taylor's series expansion, made to the rank marginal likelihood. For details of the approximation see Pettitt [2].

An observation is said to be right-censored if we can only observe Y_j^* with $Y_j^* \leq Y_j$. We rank censored and uncensored observations as follows. Suppose we can observe Y_j, for $j = 1, 2, \ldots, n$, directly but Y_j^*, for $j = n+1, n+2, \ldots, q$; $n \leq q$, are censored on the right. We define the rank r_j of Y_j, for $j = 1, 2, \ldots, n$, in the usual way; r_j equals i if and only if Y_j is the ith smallest amongst the $Y_1, Y_2, \ldots, Y_n$. The right-censored Y_j^*, for $j = n+1, n+2, \ldots, q$ has rank r_j if and only if Y_j^* lies in the interval $[Y_{(r_j)}, Y_{(r_j+1)}]$, with $Y_0 = -\infty$, $Y_{(n+1)} = +\infty$ and $Y_{(1)} < \ldots < Y_{(n)}$ the ordered Y_j, for $j = 1, 2, \ldots, n$.

The distribution of the Y's is assumed to be of the following form. Let $F_L(y) = e^y/(1+e^y)$, the logistic distribution function, and consider the distribution function $F_\gamma(y)$ defined by $1 - F_\gamma = [1 - F_L(y)]^{1/\gamma}$. This distribution function can be thought of as either the distribution function of the minimum, $X_{1,\gamma}$, of a random sample of size γ^{-1} from the logistic distribution, or as the $F_\gamma(y - \log\gamma)$ being the distribution function of a random variable having the F-distribution with 2 and $2\gamma^{-1}$ degrees of freedom. This family of generalized logistic distribution functions $[F_\gamma(.); 0 \leq \gamma < \infty]$ naturally links the symmetric logistic distribution ($\gamma = 1$) with the skew extreme value distribution ($\lim \gamma \to 0$) and with the limiting negative exponential distribution ($\lim \gamma \to \infty$). For this family explicit results are available for right-censored data. See Pettitt [3] for details.

Let l_R denote the logarithm of the rank marginal likelihood of the observations and define the $q \times 1$ vector a by $a = l_R^{'}(\theta = 0)$, and let the q by q diagonal matrix B and q by q symmetric matrix A be given by $B - A = -l_R^{''}(\theta = 0)$. Then various statistics can be found from the analysis.

(a) The score statistic $X^T a$. This statistic is used to test the hypothesis $H_0 : \beta = 0$ (see (e)).
(b) The estimated variance-covariance matrix of the score statistic in (a).
(c) The estimate $\hat{\beta}_R = MX^T a$.
(d) The estimated variance-covariance matrix $M = (X^T(B-A)X)^{-1}$ of the estimate $\hat{\beta}_R$.
(e) The χ^2 statistic $Q = \hat{\beta}_R M^{-1} \hat{\beta}_r = a^T X(X^T(B-A)X)^{-1} X^T a$, used to test $H_0 : \beta = 0$. Under H_0, Q has an approximate χ^2 distribution with p degrees of freedom.
(f) The standard errors $M_{ii}^{\frac{1}{2}}$ of the estimates given in (c).

(g) Approximate z-statistics, i.e., $Z_i = \hat{\beta}_{R_i}/s.e.(\hat{\beta}_{R_i})$ for testing $H_0 : \beta_i = 0$. For $i = 1, 2, \ldots, n$, Z_i has an approximate $N(0, 1)$ distribution.

In many situations, more than one sample of observations will be available. In this case we assume the model,

$$h_k(Y_k) = X_k^T\beta + e_k, \quad k = 1, 2, \ldots, \mathrm{NS},$$

where NS is the number of samples. In an obvious manner, Y_k and X_k are the vector of observations and the design matrix for the kth sample respectively. Note that the arbitrary transformation h_k can be assumed different for each sample since observations are ranked within the sample.

The earlier analysis can be extended to give a combined estimate of β as $\hat{\beta} = Dd$, where

$$D^{-1} = \sum_{k=1}^{\mathrm{NS}} X^T(B_k - A_k)X_k$$

and

$$d = \sum_{k=1}^{\mathrm{NS}} X_k^T a_k$$

with a_k, B_k and A_k defined as a, B and A above but for the kth sample.

The remaining statistics are calculated as for the one sample case.

4 References

[1] Kalbfleisch J D and Prentice R L (1980) *The Statistical Analysis of Failure Time Data* Wiley

[2] Pettitt A N (1982) Inference for the linear model using a likelihood based on ranks *J. Roy. Statist. Soc. Ser. B* **44** 234–243

[3] Pettitt A N (1983) Approximate methods using ranks for regression with censored data *Biometrika* **70** 121–132

5 Parameters

1: NS — INTEGER *Input*

On entry: the number of samples.

Constraint: NS ≥ 1.

2: NV(NS) — INTEGER array *Input*

On entry: the number of observations in the ith sample, for $i = 1, 2, \ldots$,NS.

Constraint: NV(i) ≥ 1, for $i = 1, 2, \ldots$,NS.

3: NSUM — INTEGER *Input*

On entry: the total number of observations.

Constraint: $\mathrm{NSUM} = \sum_{i=1}^{\mathrm{NS}} \mathrm{NV}(i)$.

4: Y(NSUM) — ***real*** array *Input*

On entry: the observations in each sample. Specifically, $\mathrm{Y}\left(\sum_{k=1}^{i-1} \mathrm{NV}(k) + j\right)$ must contain the jth observation in the ith sample.

5: IP — INTEGER *Input*

On entry: the number of parameters to be fitted.

Constraint: IP ≥ 1.

6: X(NX,IP) — ***real*** array *Input*

On entry: the design matrices for each sample. Specifically, $X\left(\sum_{k=1}^{i-1} NV(k) + j, l\right)$ must contain the value of the lth explanatory variable for the jth observations in the ith sample.

Constraint: X must not contain a column with all elements equal.

7: NX — INTEGER *Input*

On entry: the first dimension of the array X as declared in the (sub)program from which G08RBF is called.

Constraint: NX $\geq$ NSUM.

8: ICEN(NSUM) — INTEGER array *Input*

On entry: defines the censoring variable for the observations in Y as follows:

ICEN(i) $= 0$ if Y(i) is uncensored.

ICEN(i) $= 1$ if Y(i) is censored.

Constraint: ICEN(i) $= 0$ or 1, for $i = 1, 2, \ldots,$NSUM.

9: GAMMA — ***real*** *Input*

On entry: the value of the parameter defining the generalized logistic distribution. For GAMMA $\leq$ 0.0001, the limiting extreme value distribution is assumed.

Constraint: GAMMA > 0.0.

10: NMAX — INTEGER *Input*

On entry: the value of the largest sample size.

Constraint: NMAX $= \max_{1 \leq i \leq NS}(NV(i))$ and NMAX $>$ IP.

11: TOL — ***real*** *Input*

On entry: the tolerance for judging whether two observations are tied. Thus, observations Y_i and Y_j are adjudged to be tied if $|Y_i - Y_j| <$ TOL.

Constraint: TOL > 0.0.

12: PARVAR(NPVAR,IP) — ***real*** array *Output*

On exit: the variance-covariance matrices of the score statistics and the parameter estimates; the former being stored in the upper triangle and the latter in the lower triangle. Thus for $1 \leq i \leq j \leq$ IP, PARVAR(i, j) contains an estimate of the covariance between the ith and jth score statistics. For $1 \leq j \leq i \leq$ IP-1, PARVAR($i + 1, j$) contains an estimate of the covariance between the ith and jth parameter estimates.

13: NPVAR — INTEGER *Input*

On entry: the first dimension of the array PARVAR as declared in the (sub)program from which G08RBF is called.

Constraint: NPVAR $\geq$ IP $+ 1$.

14: IRANK(NMAX) — INTEGER array *Output*

On exit: for the one sample case, IRANK contains the ranks of the observations.

15: ZIN(NMAX) — ***real*** array *Output*

On exit: for the one sample case, ZIN contains the expected values of the function g(.) of the order statistics.

16: ETA(NMAX) — ***real*** array *Output*

On exit: for the one sample case, ETA contains the expected values of the function $g'(.)$ of the order statistics.

17: VAPVEC(NMAX∗(NMAX+1)/2) — ***real*** array *Output*

On exit: for the one sample case, VAPVEC contains the upper triangle of the variance-covariance matrix of the function g(.) of the order statistics stored column-wise.

18: PAREST(4∗IP+1) — ***real*** array *Output*

On exit: the statistics calculated by the routine as follows. The first IP components of PAREST contain the score statistics. The next IP elements contain the parameter estimates. PAREST(2× IP+1) contains the value of the χ^2 statistic. The next IP elements of PAREST contain the standard errors of the parameter estimates. Finally, the remaining IP elements of PAREST contain the z-statistics.

19: WORK(LWORK) — ***real*** array *Workspace*

20: LWORK — INTEGER *Input*

On entry: the dimension of the array WORK as declared in the (sub)program from which G08RBF is called.

Constraint: LWORK $\geq$ NMAX $\times$ (IP+1).

21: IWA(4∗NMAX) — INTEGER array *Workspace*

22: IFAIL — INTEGER *Input/Output*

On entry: IFAIL must be set to 0, −1 or 1. For users not familiar with this parameter (described in Chapter P01) the recommended value is 0.

On exit: IFAIL = 0 unless the routine detects an error (see Section 6).

6 Error Indicators and Warnings

If on entry IFAIL = 0 or −1, explanatory error messages are output on the current error message unit (as defined by X04AAF).

Errors detected by the routine:

IFAIL = 1

On entry, NS < 1,
or TOL ≤ 0.0,
or NMAX $\leq$ IP,
or NPVAR $<$ IP + 1,
or NX $<$ NSUM,
or NMAX $\neq \max_{1\leq i\leq \mathrm{NS}}(\mathrm{NV}(i))$,
or NV(i) ≤ 0 for some i, $i = 1, 2, \ldots, \mathrm{NS}$,
or NSUM $\neq \sum_{i=1}^{\mathrm{NS}} \mathrm{NV}(i)$,
or IP < 1,
or GAMMA < 0.0,
or LWORK $<$ NMAX $\times$ (IP + 1).

IFAIL = 2

On entry, ICEN(i) $\neq 0$ or 1 for some $1 \leq i \leq$ NSUM.

IFAIL = 3

On entry, all the observations are adjudged to be tied. The user is advised to check the value supplied for TOL.

IFAIL = 4

The matrix $X^T(B-A)X$ is either ill-conditioned or not positive-definite. This error should only occur with extreme rankings of the data.

IFAIL = 5

On entry, at least one column of the matrix X has all its elements equal.

7 Accuracy

The computations are believed to be stable.

8 Further Comments

The time taken by the routine depends on the number of samples, the total number of observations and the number of parameters fitted.

In extreme cases the parameter estimates for certain models can be infinite, although this is unlikely to occur in practice. See Pettitt [2] for further details.

9 Example

A program to fit a regression model to a single sample of 40 observations using just one explanatory variable.

9.1 Program Text

Note. The listing of the example program presented below uses bold italicised terms to denote precision-dependent details. Please read the Users' Note for your implementation to check the interpretation of these terms. As explained in the Essential Introduction to this manual, the results produced may not be identical for all implementations.

```
*     G08RBF Example Program Text
*     Mark 14 Revised.  NAG Copyright 1989.
*     .. Parameters ..
      INTEGER          NIN, NOUT
      PARAMETER        (NIN=5,NOUT=6)
      INTEGER          NSMAX, NX, NMXMAX, NSMMAX, IPMAX, NPVAR, LWORK
      PARAMETER        (NSMAX=5,NX=100,NMXMAX=NX,NSMMAX=NX,IPMAX=6,
     +                 NPVAR=IPMAX+1,LWORK=NMXMAX*(IPMAX+1))
*     .. Local Scalars ..
      real             GAMMA, TOL
      INTEGER          I, IFAIL, IP, J, NMAX, NS, NSUM
*     .. Local Arrays ..
      real             ETA(NMXMAX), PAREST(4*IPMAX+1),
     +                 PARVAR(NPVAR,IPMAX), VAPVEC(NMXMAX*(NMXMAX+1)/2),
     +                 WORK(LWORK), X(NX,IPMAX), Y(NSMMAX), ZIN(NMXMAX)
      INTEGER          ICEN(NSMMAX), IRANK(NMXMAX), IWA(4*NMXMAX),
     +                 NV(NSMAX)
*     .. External Subroutines ..
      EXTERNAL         G08RBF
*     .. Intrinsic Functions ..
      INTRINSIC        MAX
*     .. Executable Statements ..
      WRITE (NOUT,*) 'G08RBF Example Program Results'
```

```
*     Skip heading in data file
      READ (NIN,*)
*     Read number of samples, number of parameters to be fitted,
*     distribution power parameter and tolerance criterion for ties.
      READ (NIN,*) NS, IP, GAMMA, TOL
      WRITE (NOUT,*)
      IF (NS.GT.0 .AND. NS.LE.NSMAX .AND. IP.GT.0 .AND. IP.LE.IPMAX)
     +    THEN
         WRITE (NOUT,99999) 'Number of samples =', NS
         WRITE (NOUT,99999) 'Number of parameters fitted =', IP
         WRITE (NOUT,99998) 'Distribution power parameter =', GAMMA
         WRITE (NOUT,99998) 'Tolerance for ties =', TOL
*        Read the number of observations in each sample
         READ (NIN,*) (NV(I),I=1,NS)
         NMAX = 0
         NSUM = 0
         DO 20 I = 1, NS
            NSUM = NSUM + NV(I)
            NMAX = MAX(NMAX,NV(I))
   20    CONTINUE
         IF (NMAX.GT.0 .AND. NMAX.LE.NMXMAX .AND. NSUM.GT.0 .AND.
     +       NSUM.LE.NSMMAX) THEN
*           Read in observations, design matrix and censoring variable
            READ (NIN,*) (Y(I),(X(I,J),J=1,IP),ICEN(I),I=1,NSUM)
            IFAIL = 0
*
            CALL G08RBF(NS,NV,NSUM,Y,IP,X,NX,ICEN,GAMMA,NMAX,TOL,PARVAR,
     +                  NPVAR,IRANK,ZIN,ETA,VAPVEC,PAREST,WORK,LWORK,
     +                  IWA,IFAIL)
*
            WRITE (NOUT,*)
            WRITE (NOUT,*) 'Score statistic'
            WRITE (NOUT,99997) (PAREST(I),I=1,IP)
            WRITE (NOUT,*)
            WRITE (NOUT,*) 'Covariance matrix of score statistic'
            DO 40 J = 1, IP
               WRITE (NOUT,99997) (PARVAR(I,J),I=1,J)
   40       CONTINUE
            WRITE (NOUT,*)
            WRITE (NOUT,*) 'Parameter estimates'
            WRITE (NOUT,99997) (PAREST(IP+I),I=1,IP)
            WRITE (NOUT,*)
            WRITE (NOUT,*) 'Covariance matrix of parameter estimates'
            DO 60 I = 1, IP
               WRITE (NOUT,99997) (PARVAR(I+1,J),J=1,I)
   60       CONTINUE
            WRITE (NOUT,*)
            WRITE (NOUT,99996) 'Chi-squared statistic =',
     +        PAREST(2*IP+1), ' with', IP, ' d.f.'
            WRITE (NOUT,*)
            WRITE (NOUT,*) 'Standard errors of estimates and'
            WRITE (NOUT,*) 'approximate z-statistics'
            WRITE (NOUT,99995) (PAREST(2*IP+1+I),PAREST(3*IP+1+I),I=1,
     +        IP)
         END IF
      END IF
      STOP
*
```

```
99999 FORMAT (1X,A,I2)
99998 FORMAT (1X,A,F10.5)
99997 FORMAT (1X,F9.3)
99996 FORMAT (1X,A,F9.3,A,I2,A)
99995 FORMAT (1X,F9.3,F14.3)
      END
```

9.2 Program Data

```
G08RBF Example Program Data
 1 1 0.00001 0.00001
 40
 143.0 0.0 0 164.0 0.0 0 188.0 0.0 0 188.0 0.0 0 190.0 0.0 0
 192.0 0.0 0 206.0 0.0 0 209.0 0.0 0 213.0 0.0 0 216.0 0.0 0
 220.0 0.0 0 227.0 0.0 0 230.0 0.0 0 234.0 0.0 0 246.0 0.0 0
 265.0 0.0 0 304.0 0.0 0 216.0 0.0 1 244.0 0.0 1 142.0 1.0 0
 156.0 1.0 0 163.0 1.0 0 198.0 1.0 0 205.0 1.0 0 232.0 1.0 0
 232.0 1.0 0 233.0 1.0 0 233.0 1.0 0 233.0 1.0 0 233.0 1.0 0
 239.0 1.0 0 240.0 1.0 0 261.0 1.0 0 280.0 1.0 0 280.0 1.0 0
 296.0 1.0 0 296.0 1.0 0 323.0 1.0 0 204.0 1.0 1 344.0 1.0 1
```

9.3 Program Results

```
 G08RBF Example Program Results

 Number of samples = 1
 Number of parameters fitted = 1
 Distribution power parameter =   0.00001
 Tolerance for ties =   0.00001

 Score statistic
     4.584

 Covariance matrix of score statistic
     7.653

 Parameter estimates
     0.599

 Covariance matrix of parameter estimates
     0.131

 Chi-squared statistic =    2.746 with 1 d.f.

 Standard errors of estimates and
 approximate z-statistics
     0.361         1.657
```